Proceedings of ICAD2011
The Sixth International Conference on Axiomatic Design
Daejeon – March 30-31, 2011
ICAD-2011-24

SCHEDULING PRE-ADMISSION TESTING
A CASE STUDY ON REDUCING COMPLEXITY IN HEALTHCARE DELIVERY

Hilario L. Oh
ohlarry@mit.edu
MIT Park Center for Complex Systems
Massachusetts Institute of Technology
Cambridge, MA 02139 USA

ABSTRACT

Healthcare delivery in today's hospitals is a complex task. A major source of complexity comes from the coupling of the various operations in the hospital. Because of coupling, any attempt to optimize one function would be at the expense of other functions. The consequence is a complex, sub-optimal system. Thus, to arrive at an optimized healthcare delivery system, we need to identify and resolve the couplings present so that the complexity of the system may be reduced and the system optimized.

In this paper, we use the scheduling of Pre-Admission Testing (PAT) for surgical patients to illustrate the reduction of complexity that leads to optimized healthcare delivery. We first review the PAT process. We next analyze the process to reveal the three functional requirements of PAT: increase throughput, reduce work-in-progress (WIP) and deliver results on-time. We then show that the corresponding scheduling algorithms, longest-process-time first, shortest-process-time first and least-slack-time first, that are traditionally used to optimize each of the three functional requirements in fact couple them together. Consequently, the PAT scheduler becomes complex and sub-optimal. Finally, we introduce a two-stage PAT scheduler to resolve the couplings among the three functional requirements. The resolution of couplings simplifies the system and allows each scheduling algorithm to independently optimize the corresponding functional requirement. The result is a dramatic reduction in WIP and improvement of on-time delivery.

Keywords: Two-stage scheduler, pre-admission testing.

1 BACKGROUND

Healthcare delivery in today's hospitals is a complex task. A major source of complexity comes from the coupling of multiple operations in the hospital. The coupling causes the function of units like emergency departments (EDs), inpatient units, operating rooms (ORs), the central sterilization department (CSD) and intensive care units (ICUs) to be interdependent. Inefficient boarding in an inpatient unit that results in a back up of patient flow and crowding in the ED is an example of the coupling of operations and processes in a hospital. Boarding is the term used to describe the process of scheduling a patient for a procedure or otherwise entering the patient into the system of a particular unit. Another related source of complexity derives from the coupling of various processes and resources required to complete a task within a unit. In the OR, providing a greater degree of accessibility to surgeons or patients at the expense of capacity utilization or productivity is an example of coupling of processes or resources within a unit. Because the coupling of operations and processes brought about interdependency among the functions of the units of the healthcare delivery system, it is apparent that these operations and processes need to be managed as a whole, i.e., as a 'system'. One approach to manage healthcare delivery as a system is to use Axiomatic Design as exemplified by Kolb *et al.* [2007] and Peck *et al.* [2009].

To manage a system of operations, we first define the functional requirements FRs of the operations that comprise the system. Next we choose the design parameters DPs that render the FRs mutually exclusive, free of couplings and interdependencies. The latter step is critical because the degree of coupling in a design is determined primarily by the choice of DPs and not the physics governing the design. We illustrate this point with the faucet example.

In the discourse of Axiomatic Design, the faucet example has been used over and over again to give an intuitive explanation of coupling. Here, we shall give a mathematical explanation as follows. The FRs are:

FR_1: deliver water at a certain flow rate Q;
FR_2: deliver water at a certain temperature T.

Assuming that we are mixing hot and cold water, we may express the two FRs (Q, T) in terms of the hot and cold water flow rates (Q_h, Q_c) and temperatures (T_h, T_c) as follows.

Mass conservation: $Q = Q_h + Q_c$ (1)
Energy conservation: $TQ = T_hQ_h + T_cQ_c$

So that
$$T = \frac{T_hQ_h + T_cQ_c}{Q_h + Q_c} \quad (2)$$

Equations (1) and (2) express the FRs in mathematical terms of mixing hot and cold waters. We now decide on the DPs that affect the FRs. We may choose (Q_h, Q_c) to be (DP_1, DP_2) respectively. In this case, in Faucet Design No. 1, the FRs are related to the DPs through the matrix Equation (3) derived from Equations (1) and (2) as follows.

$$\begin{Bmatrix} Q \\ T \end{Bmatrix} = \begin{bmatrix} 1 & 1 \\ \dfrac{T_h}{Q_h + Q_c} & \dfrac{T_c}{Q_h + Q_c} \end{bmatrix} \begin{Bmatrix} Q_h \\ Q_c \end{Bmatrix} \quad (3)$$

Thus in making this choice of $(DP_1, DP_2) = (Q_h, Q_c)$, we have coupled the two FRs together. Namely, a choice of (Q_h, Q_c) pair to achieve a certain flow rate Q also affects the temperature T and vice versa. Consequently, iterative trial and error of the (Q_h, Q_c) pair are needed to converge to a desired set of (Q, T). In short, there is interdependency between FR_1, the flow rate Q and FR_2, the temperature T.

We can make an alternative choice of DPs to break up the interdependency between FR_1 and FR_2. We choose the sum (Q_h+Q_c) as DP_1 and the ratio (Q_h/Q_c) as DP_2. As derived from Equations (1) and (2), this choice, Faucet Design No. 2, yields the matrix Equation (4) as follows

$$\begin{Bmatrix} Q \\ T \end{Bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & \dfrac{(Q_h/Q_c)T_h + T_c}{[(Q_h/Q_c)+1](Q_h/Q_c)} \end{bmatrix} \begin{Bmatrix} Q_h + Q_c \\ (Q_h/Q_c) \end{Bmatrix} \quad (4).$$

With this choice, we render the two FRs independent of each other. That is, we can attain a flow rate Q with (Q_h+Q_c) independent of attaining a temperature T with (Q_h/Q_c). Note that the governing physics for both designs is the same. It is the choice of the DPs that determine the degree of coupling.

The above faucet design examples illustrate the approach to managing a complex system of operations: define the *right* FRs, make sure they are collectively exhaustive; then choose the DPs *right* to eliminate couplings among the FRs so that they are mutually exclusive. Once uncoupled, the FRs can be optimized independent of one another. We now apply this approach to healthcare delivery systems.

2 SYSTEM APPROACH TO HEALTH CARE DELIVERY

There is a myriad of FRs for health care delivery. The more common ones in operations and processes and their corresponding metrics are:

- raise throughput - number of jobs completed per unit time, the larger the better;
- reduce the length of stay - length of time a job spends in the system, the shorter the better;
- use resources optimally - level of capacity utilization, the higher the better;
- reduce wait time - length of time between the arrival of a job and the start of processing the job, the shorter the better;
- deliver service on time - difference between due date and completion time, the larger the better;
- deliver service reliably - consistency in delivering a job on time, the more consistent the better.

The DPs for the above FRs typically involved the following process variables:

- processing time - the length of time to process a job;
- due date - last "time" to complete the job;
- capacity - amount of resources available; e.g., number of beds available;
- slack time - job's time to due date (TDD) minus its processing time;
- completion time - time at which a job is finished

In the discussion above, two or more FRs may involve one and the same process variable. There is thus a potential for coupling among them. For example in EDs, both the throughput and length of stay are dependent on the processing time that it takes to diagnose and treat a patient. Obviously, the two FRs cannot be independently satisfied by one process variable. Hence, they are coupled. Similarly, both on time delivery and reliable delivery could be coupled because they involve the same process variable TDD, the time to due date. In operations research, tools and techniques are available which are used to optimize the above FRs, e.g., minimize the length of stay or maximize throughput. These tools typically do not take coupling among the FRs into account, addressing instead a specific FR at the expense of other FRs. The consequence is a sub-optimal system. To ensure that we attain an optimal system, we must first resolve the coupling among the FRs before using the tools to optimize them. We illustrate this with a case study, scheduling the PAT of surgical patients. This case study was the basis for a patent filed with the US patent office [Oh, 2009].

3 SCHEDULING PRE-ADMISSION TESTING (PAT) OF SURGICAL PATIENTS

3.1 THE PAT PROCESS

The PAT process is a critical step in healthcare delivery, as its goal is to ensure that a patient's pre-operative conditions will not affect his scheduled surgical procedure. Toward this goal, the medical staff has developed requirements for pre-procedure testing based on the patient's pre-operative condition and surgical procedure scheduled. For example, one requirement is that all patients age >74 must have an electro-cardiogram and complete blood tests regardless of the planned surgical procedure. Another requirement is that a patient scheduled for a total hip replacement must have complete blood tests and blood typing and cross matching regardless of age and pre-operative condition. Based on the requirements developed, the PAT process screens the patient for pre-procedure testing needs and gathers results of the needed tests to clear the patient for the planned surgical procedure.

Figure 1 shows the timeline that depicts the events completed during the PAT process. The PAT process starts when the patient is boarded for a surgical procedure. It ends when the patient is transferred to the OR for the surgical procedure. The interceding time period is referred to as "board-to-surgery" time.

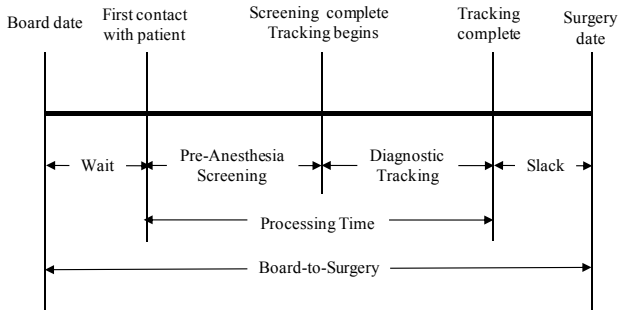

Figure 1. Timeline showing events in the PAT process.

After a patient is boarded, there is a wait period before the PAT nurse makes the first contact with the patient. The wait exists because the PAT nurse is most likely busy with another patient. Once contacted, the PAT nurse interviews the patient for his or her health history and status to determine the need for pre-procedure testing. If the determination calls for tests for which the patient already has lab results, the PAT nurse simply collects them. If existing results are not available, the PAT nurse then orders the tests. The PAT nurse also identifies patients with learning needs and refers the patient to an education program e.g., a class on total hip replacement. The time between when the PAT nurse first contacts the patient and when she completes the interview and obtains all of the pre-procedure testing needs is called the pre-anesthesia screening.

After the pre-anesthesia screening, the PAT nurse starts tracking and gathering the results of the pre-procedure tests that were called for. This involves tracking and gathering existing lab results from the patient's primary physician and new results from additional tests that were ordered. The time spent in tracking and gathering the test results is called diagnostic tracking. Time spent in tracking varies according to how many and what type of tests are being tracked. The sum total of screening and tracking time constitutes the processing time of the PAT process. Ideally, the PAT process should be completed ahead of the due date, the day of surgery, for two reasons: (1) to provide a buffer for variability in processing time; and (2) to be able to charge for pre-procedural tests done. Usually, the insurance companies will not pay for pre-procedural tests completed within three days of the surgical procedure. They treat these tests as part of the surgical procedure itself.

If the tracking process is not complete by the due date, the surgical procedure must be rescheduled. Rescheduling a surgical procedure wastes a tremendous amount of resources. Typically, the surgeons and doctors that were to perform the procedures do not have enough notice to fill the time slot of the patient, thereby resulting in a loss of revenue for that time slot. Moreover, many of the pre-procedure tests will have to be repeated, as test results are normally valid only for 30 days or less. Additionally, the PAT nurse will have to re-track the patient. Thus, it is imperative that the tracking be completed by the due date. The period between the end of the tracking and the due date is referred to as slack time.

3.2 Functional Requirements and Design Parameters of PAT

Given a stream of patients for PAT processing, which hereafter we refer to as jobs, what are the functional requirements (FRs) and design parameters (DPs) for processing these jobs? An FR that a PAT process traditionally aims for is

FR_1: raise the throughput.

Throughput is the number of jobs completed per unit time. For a stream of n PAT jobs assigned to m PAT nurses, the ith nurse, $i=1, 2 \ldots m$, is assigned a subset of n jobs for a certain total time T_i. The maximum total time, T_{max} of the set T_i, $i=1, 2 \ldots m$ is the time that it takes to complete the n jobs by the m nurses as a group. Therefore the throughput of the m nurses as a group is equal to n/T_{max}. To raise the throughput is to find an assignment of jobs to the nurses that reduce T_{ma}

To illustrate, consider the assignment of eight jobs, (n = 8) to three nurses, (m = 3) as shown in Figure 2.

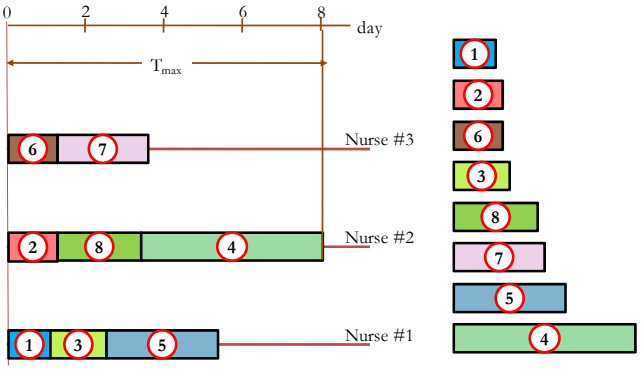

Figure 2. SPT assignment of eight jobs to three nurses.

The jobs are represented by rectangles with the processing time in days indicated by the length of the rectangle. One scheduling algorithm, Shortest Processing Time (SPT) first, is to sort the jobs in order of increasing processing times. Whenever a nurse is freed, the job with the shortest processing time at that instant is assigned to her for processing. With this scheduling algorithm, nurse #2 ends up with two jobs for a total time of 8 days, the maximum of the three. Therefore, the throughput of the three nurses is 8 jobs in 8 days, or 1.0 job per day

For contrast, consider an alternative scheduling algorithm, the Longest Processing Time (LPT) first. The LPT sorts the jobs in order of decreasing processing times. Whenever a nurse is freed, the job with the longest processing time at that instant is assigned to her for processing. By scheduling the longest jobs first, this algorithm ensures that no one large job "sticks out" at the end of the schedule to dramatically lengthen the completion time of the jobs. Consequently, the three nurses are assigned the eight jobs for about the same total time of 6 days each, see Figure 3.

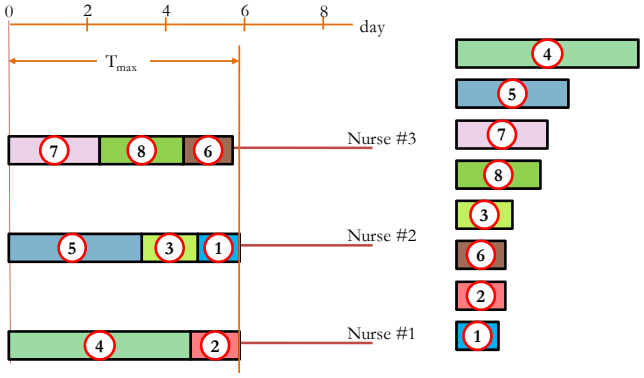

Figure 3. LPT assignment of eight jobs to three nurses.

In other words, LPT utilizes the three nurses fully, with none of them idling. The throughput for the group is 8 jobs in 6 days; or 1.333 jobs per day, a 33.3% increase over the SPT algorithm. Thus, to raise throughput, the DP for FR_1 is LPT.

DP_1: LPT, process job with the longest processing time first.

Another FR that PAT process traditionally aims for is

FR_2: reduce holding of jobs (WIP) in the system.

WIP is essentially an intermediate storage of job in a system due to lack of capacity. Thus to reduce WIP, we create a larger capacity. The added capacity and related flexibility enable us to absorb short term fluctuations.

Consider two consecutive jobs with processing time P_1 and P_2 being assigned to a PAT nurse for processing. The first job spends P_1 time in the system completing the process. The second job spends P_1 time in the system waiting for the first job to complete and then spends P_2 time completing the process. The total time spent by the two jobs in the system is $(2P_1 + P_2)$. In general, the total time spent by n jobs in a system is $nP_1 + (n\text{-}1)P_2 + (n\text{-}2)P_3 + \ldots + 2P_{n-1} + P_n$. It is the least when the jobs are ordered in increasing order: $P_1 < P_2 < \ldots < P_{n-1} < P_n$. Therefore to reduce the holding of jobs in the system, which is directly proportional to the total time spent in the system by these jobs, we process the jobs with the shortest processing time first. That is, the DP for FR_2 is SPT.

DP_2: SPT, process job with the shortest processing time first

Finally, a third and the primary FR for PAT process is

FR_3: ensure jobs are processed and delivered on time.

Since the lateness of a job is the measure of performance for FR_3, the time difference between the due date and the completion time is therefore the measure of urgency. The DP widely used for FR_3 is EDD, process jobs with the earliest due date first.

DP_3: EDD, process jobs with the earliest due date first.

Summarizing the discussion above, we have the functional requirements and their design parameters of PAT as follows.

FR_1: raise throughput;
DP_1: LPT, process jobs with longest processing time first.
FR_2: reduce WIP;
DP_2: SPT, process jobs with shortest processing time first.
FR_3: deliver on time;
DP_3: EDD, process jobs with the earliest due date first.

Each of the DPs above is a well known heuristic in operations research used for optimizing the corresponding FR [Leung, 2004]. However, each heuristic does not take into account the couplings among the FRs; addressing instead only a specific FR at the expense of other FRs. Consequently, the system is sub-optimal. In the next section, we take a system approach by considering the three FRs together as a whole, determining where the couplings among them are; resolving the couplings and optimizing the FRs independent of other FRs to arrive at an optimum system.

3.3 Identifying Coupling of Functional Requirements in PAT

Raise throughput, FR_1 and reduce WIP, FR_2 are coupled because both are dependent on one and the same DP: the ordering and sequencing of processing time. LPT, the best order and sequence of processing time for FR_1 is in fact the worst for FR_2; while SPT, the best order and sequence of processing time for FR_2 is in fact the worst for FR_1. This coupling of FR_1 and FR_2 is analogous to that of Faucet Design No. 1. It maybe symbolically represented as

$$\begin{Bmatrix} \text{Raise throughput} \\ \\ \text{Reduce WIP} \end{Bmatrix} = \begin{bmatrix} X & X \\ X & X \end{bmatrix} \begin{Bmatrix} \text{LPT} \\ \\ \text{SPT} \end{Bmatrix} \qquad (5)$$

In the above and hereafter, an element "X" in the matrix indicates that a DP has an effect on an FR; and an "O" has no effect. So a diagonal element X_{ii} indicates that DP_i has an effect on FR_i. An off-diagonal element "X_{ij}" indicates that DP_j, which has an effect on FR_j, also has an effect on FR_i . In other words, DP_j couples FR_i to FR_j.

Deliver on time, FR_3 is dependent not only on TDD, time to due date but also on processing time. This is because a job with an early due date can still be completed and delivered on time if it has a short processing time. Thus, FR_3 is affected not only by DP_3(=EDD) but by DP_1(=LPT) and DP_2(=SPT) as well:

$$\{\text{Deliver on time}\} = \begin{bmatrix} X & X & X \end{bmatrix} \begin{Bmatrix} \text{LPT} \\ \\ \text{SPT} \\ \\ \text{EDD} \end{Bmatrix} \qquad (6)$$

Combining Equations (5) and (6), we have

$$\begin{Bmatrix} \text{Raise throughput} \\ \text{Reduce WIP} \\ \text{Deliver on time} \end{Bmatrix} = \begin{bmatrix} X & X & O \\ X & X & O \\ X & X & X \end{bmatrix} \begin{Bmatrix} \text{LPT} \\ \text{SPT} \\ \text{EDD} \end{Bmatrix} \quad (7)$$

The matrix in Equation (7) summarizes the couplings among the FRs in PAT as indicated by the presence of "X" in the off-diagonal elements:

- "X" in row 1, column 2 indicates coupling of FR_1 with FR_2 through SPT;
- "X" in row 2, column 1 indicates coupling of FR_2 with FR_1 through LPT;
- "X" in row 3, columns 1 and 2 indicate coupling of FR_3 with FR_1 and FR_2 through LPT and SPT.

These sources of coupling need to be resolved for the system to be optimal.

3.4 Resolving Coupling of Functional Requirements in PAT

3.4.1 Adopting a New Process Variable to Uncouple FR_3

We resolve the coupling of FR_3 with FR_1 and FR_2 by choosing an alternative DP_3 involving a different process variable, the slack. As illustrated in Figure 4, given the date that a job is ready to start, the urgency of the job is dependent not only on TDD, the time to due date; but also on the processing time PT (= screening + tracking) it takes to complete the job.

For example in Figure 4, even though due date of Job #1 is earlier than that of Job#2, Job#1 is actually less urgent

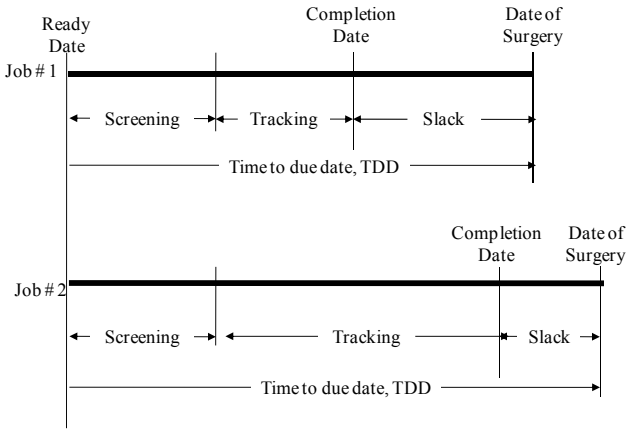

Figure 4. Comparison of urgency between two jobs.

since its processing time is shorter. In other words due date, the variable widely used in PAT as a measure for urgency, is only a partial measure. The other part is the processing time to complete the job. Thus, slack is the more relevant measure since it takes both TDD and processing time PT into account:

Slack = (TDD – PT).

Furthermore, slack ensures consistency in on-time delivery because it serves as a buffer to absorb variability in processing time. A shorter slack indicates not only higher urgency to deliver on time but also a smaller margin for consistency in on-time delivery. Thus the logical DP for FR_3 is a scheduling algorithm of jobs based on their slack times.

DP_3: LST, process jobs with the least slack time first.

With the choice of LST as DP_3, FR_3 is uncoupled from FR_1 and FR_2. FR_3 becomes solely dependent DP_3, completely independent of LPT and SPT as shown below

$$\begin{Bmatrix} \text{Raise throughput} \\ \text{Reduce WIP} \\ \text{Deliver on time} \end{Bmatrix} = \begin{bmatrix} X & X & O \\ X & X & O \\ O & O & X \end{bmatrix} \begin{Bmatrix} \text{LPT} \\ \text{SPT} \\ \text{LST} \end{Bmatrix} \quad (8)$$

3.4.2 Introducing a Two-stage Scheduler to Resolve the Couplings of FR_1 and FR_2

We now describe a two-stage scheduler that resolves the remaining coupling between FR_1 and FR_2 as indicated in Equation (8). FR_1 and FR_2 are coupled because they are dependent essentially on one and the same DP: the ordering and sequencing of the processing time, PT. Furthermore, LPT and SPT are in contradiction: LPT demands the sequencing of processing time in decreasing order while SPT demands sequencing in increasing order. To resolve the coupling and contradiction, we split the traditional PAT process as shown in Figure 5 into two stages, screening and tracking, as shown in Figure 6.

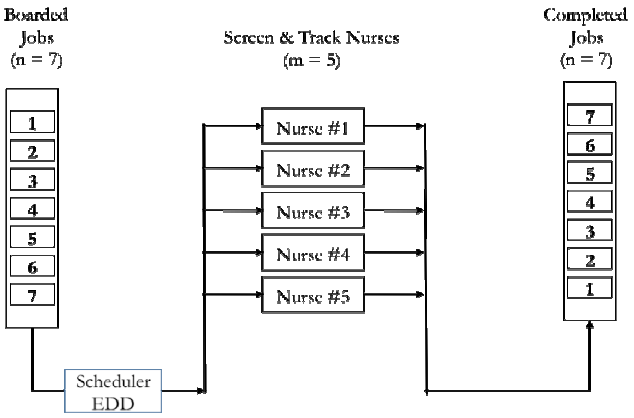

Figure 5. Traditional single stage PAT processing.

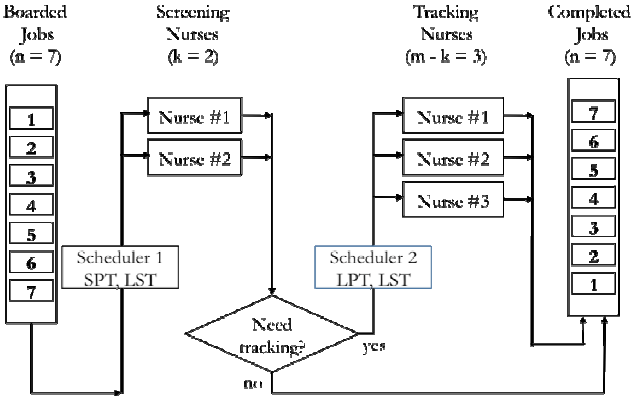

Figure 6. Proposed two-stage PAT processing.

In the traditional PAT process, n (= 7) jobs are assigned to m (= 5) nurses in parallel, with each nurse performing both screening and tracking tasks. In the proposed two-stage setup, k (= 2) of the m nurses in parallel are assigned to the first stage to perform the screening task only. Once screened, jobs that need no pre-procedure testing thus no tracking immediately exit the system. The remaining jobs that need pre-procedure tests are then sent to the second stage for tracking by the remaining $m - k$ (= 3) nurses in parallel. The apportionment of m nurses to k screening and (m-k) tracking is done in proportion to the workloads at the two stages.

The primary intent of the screening stage is to capture two categories of jobs: (1) jobs that need no pre-procedure testing and (2) jobs that require more than three pre-procedure tests. Jobs in category 1 need no tracking and therefore may exit the system immediately, reducing the WIP. Jobs in category 2 have short slack time and thus need to be sent quickly to the tracking stage for immediate tracking to ensure on-time delivery. This intent of the screening stage is represented by Equation (9) below.

$$\begin{Bmatrix} \text{Reduce WIP} \\ \\ \text{Deliver on time} \end{Bmatrix} = \begin{bmatrix} X & O \\ \\ O & X \end{bmatrix} \begin{Bmatrix} SPT \\ \\ LST \end{Bmatrix} \qquad (9)$$

Screening Stage

At the tracking stage, we raise the throughput with the algorithm LPT; and maximize the consistency in on-time delivery with the algorithm LST.

$$\begin{Bmatrix} \text{Raise throughput} \\ \\ \text{Deliver on time} \end{Bmatrix} = \begin{bmatrix} X & O \\ \\ O & X \end{bmatrix} \begin{Bmatrix} LPT \\ \\ LST \end{Bmatrix} \qquad (10)$$

Tracking Stage

Thus by splitting the PAT process into two stages as represented by Equations (9) and (10), we resolve the coupling and contradiction of LPT and SPT originally present in the traditional single stage PAT process Equation (7).

3.5 POLICY FOR DISPATCHING JOBS

Historical data suggests that jobs that require screening but no pre-procedure testing involve patients with age < 50 that need only outpatient care. These jobs comprise 25% of the total jobs. They require on the average 1.33 days to screen, i.e., PT = 1.33 day. The corresponding slack time is equal to (TDD - PT) is (TDD-1.33). Historical data also suggests that jobs that need more than three pre-procedure tests are patients that need inpatient care, who are undergoing orthopedic or vascular surgery. These jobs comprise 35% of the total jobs. They are suspected of having short slack time. Therefore we impose that the slack time for these jobs be no less than 1 day. The corresponding processing time is thus (TDD-1). For all other jobs, the average of the two estimates above is used. That is: for the processing time, (TDD + 0.33)/2; for the slack time, (TDD-0.33)/2. Note that the TDD of a job entering the screening stage is known. Thus the processing time and the slack time of the job may be estimated per Col (1) and Col (2) in Table 1 below.

Since the DP for reducing WIP is the shortest-processing-time first; and the DP for delivering service on time is the least-slack-time first, a priority value for dispatching jobs at the screening stage may be expressed as the combination of the two DPs:

$$\text{Priority value} = \frac{1}{\text{Processing Time x Slack}}$$

.

$$= \left[PT x (TDD - PT)\right]^{-1}$$

The above priority value maybe calculated per Col (3) in Table 1 below. The policy for dispatching jobs at the screening stage is to process the job with the highest priority value first.

Once a job has been screened, the type and number of pre-procedure tests needed for the job are known. The associated processing time PT may therefore be estimated from historical data. Furthermore, the TDD is known for each job. Thus, the slack time (TDD - PT) may be estimated as well. Since the DP for raising the throughput is the longest-processing-time first; and the DP for delivering service on time is the least-slack-time first, a priority value for dispatching jobs at the tracking stage may be expressed as the combination of the two DPs:

$$\text{Priority value} = \frac{\text{Processing Time}}{\text{Slack}}$$

$$= \frac{PT}{(TDD - PT)}$$

The policy for dispatching jobs at this stage is to process the job with the highest priority value first.

Table 1. Formulae for processing time, slack and priority value at the screening stage.

Category	Processing time Col(1)	Slack Col(2) = TDD – Col(1)	Priority value Col(3) =[Col(1) x Col(2)]$^{-1}$
Outpatient, age<50	1.33	TDD – 1.33	[1.33 x (TDD - 1.33)]$^{-1}$
Orthopaedic or vascular	TDD - 1	1	(TDD – 1)$^{-1}$
Otherwise	(TDD + 0.33)/2	(TDD - 0.33)/2	4 x [TDD^2 – (0.33)^2]$^{-1}$

4 DISCUSSION OF RESULTS

A two-stage PAT scheduler was launched on 3/24/2008 at a local hospital in Rochester, Michigan, USA. Two "snapshots" of the WIP were taken; one snapshot a month before and another, a month after the launch. As shown superimposed in Figure 7, each snapshot depicts a month, equals 21 working days, of PAT jobs waiting in queue to be

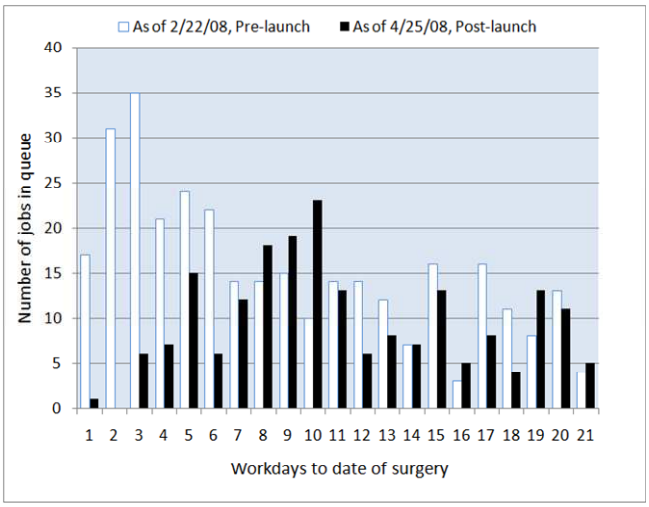

Figure 7. Snapshots of jobs in queue.

processed. The snapshot shown in "white" taken before the launch is that of the traditional single stage PAT scheduler. It shows a WIP of 321 jobs. It also shows 150 (47%) of them are within 6 workdays from due date, the day of surgery. In fact, 83 of them (25%) are within 3 workdays from due date. These are jobs that will not be paid for by the insurance companies. It represents a loss of revenue for the hospital. The occurrence of a large WIP, with the majority of the jobs close to due date is to be expected. This is because the FRs, reduced WIP and ample slack, were never defined and aimed for in the single stage scheduler. Only EDD, a partial measure of urgency was implemented. In other words, we were not *doing the right things*.

The other snapshot shown in "black" taken after the launch is that of the two-stage scheduler. It shows a WIP of 200 jobs, significantly reduced from the original WIP of 321 jobs. Furthermore, there is now ample slack with 193 of the total 200 jobs (96.5%) a distant 4 workdays away from the due date. This is the result of deliberately define the *right* set of FRs for the scheduler: reduce WIP, raise throughput and increase slack. This is followed by implementing the DPs *right*:: remove couplings among the FRs that then permits the DPs, SPT, LPT and LST, to optimize each FR independent of others. In short, we *did the right things* and *did the things right*.

5 ACKNOWLEDGMENTS

The preparation and presentation of this paper at the conference was supported and funded by The MIT Park Center for Complex Systems at Massachusetts Institute of Technology.

6 REFERENCES

[1] Kolb E., Lee T., Peck J., "Effect of coupling between emergency department and inpatient unit on overcrowding in emergency department", *IEEE Winter Simulation Conference*, December, 2007.

[2] Leung J. Y-T., *Handbook of scheduling: algorithms, models, and performance analysis*, Chapman & Hall/CRC Press, 2004.

[3] Oh H.L., "Method and system for managing operations and processes in healthcare delivery in a hospital", US Patents Pending, Application Publication #US2010/041452 A1, 2009.

[4] Peck J., Lee T., Kolb E., and Kim S., "Lessons learned by applying axiomatic design to an emergency department", *Proceeding of The Fifth International Conference on Axiomatic Design,* March 2009.